Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte

Brigitte Escofier

IRIA

CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.: 954 90 20

Rapports de Recherche

N° 82

TRAITEMENT DES QUESTIONNAIRES AVEC NON RÉPONSE, ANALYSE DES CORRESPONDANCES AVEC MARGE MODIFIÉE ET ANALYSE MULTICANONIQUE AVEC CONTRAINTE

Brigitte ESCOFFIER

Juin 1981

IRIA

CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.: 954 90 20

Rapports de Recherche

N° 82

TRAITEMENT DES QUESTIONNAIRES AVEC NON RÉPONSE, ANALYSE DES CORRESPONDANCES AVEC MARGE MODIFIÉE ET ANALYSE MULTICANONIQUE AVEC CONTRAINTE

Brigitte ESCOFFIER

Juin 1981

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTEMES ALEATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
TéL. : (99) 36 48 15
Télex : UNIRISA 95 0473 F

TRAITEMENT DES QUESTIONNAIRES AVEC NON REPONSE, ANALYSE DES CORRESPONDANCES AVEC MARGE MODIFIEE ET ANALYSE MULTICANONIQUE AVEC CONTRAINTE

Brigitte ESCOFFIER

Publication Interne n° 146 - Mai 1981
38 pages

RESUME

Pour traiter les questionnaires où manquent certaines réponses, et, d'une façon générale, les tableaux disjonctifs non complets, nous proposons des variantes de méthodes classiques d'analyse des données :

Une variante de l'analyse des correspondances où l'une des marges du tableau de donnée est remplacée par une marge imposée ; une analyse multicanonique avec contrainte linéaire sur les facteurs.

Appliquées au problème posé, ces deux variantes donnent les mêmes facteurs et permettent une analyse ayant la plupart des propriétés de l'analyse classique des tableaux disjonctifs complets. D'autres applications sont possibles, par exemple l'ajustement d'un tableau de fréquence à des marges.

SUMMARY

We propose some variations of the classical methods of data analysis for anysing the questionnaires with some missing answers, and in general, for analysing the incomplete disjunctive tables :

A variation of correspondance analysis where one of the marginal distributions of the data table is replaced by an imposed marginal distribution ; a multicanonical analysis with linear constraints on factors.

Application of the two variations to the present problem produces the same factors, and allows the analysis with most of the properties of the classical analysis of the complete disjunctive tables. Other applications one possible, for example, fitting a frequency table to the marginal distributions.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

- TABLE -

TRAITEMENT DES QUESTIONNAIRES AVEC NON REPONSE. ANALYSE DES CORRESPONDANCES AVEC MARGE MODIFIEE ET ANALYSE MULTICANONIQUE AVEC CONTRAINTE

Brigitte ESCOFFIER

I - LE PROBLEME DES QUESTIONNAIRES AVEC NON REPONSE.

I.a - Rappel sur les tableaux disjonctifs complets

L'ensemble des réponses à un questionnaire peut être codé par un tableau disjonctif complet lorsque l'on impose à tous les individus de choisir, pour chaque question, une réponse et une seule parmi celles qui sont proposées.

Ce tableau croise l'ensemble des individus de la population et l'ensemble des réponses à toutes les questions. Il comporte un 1 au croisement de la i-ième ligne et de la j-ième colonne si l'individu i a choisi la réponse j et des zéros partout ailleurs.

L'efficacité du traitement de ce type de tableau par l'analyse des correspondances -qu'on appelle souvent dans ce cas, analyse des correspondances multiples- n'est plus à démontrer. [cf. 1, 2, 12, 13].

Les résultats de cette analyse mettent bien en évidence les groupes d'individus ayant choisi des réponses semblables, et les groupes de réponses choisies généralement par les mêmes individus.

On sait aussi que cette analyse peut être définie à partir des questions elles-mêmes et non plus des réponses à ces questions. [cf. 2, 4, 15]. Nous le rappelons ici, car nous utiliserons aussi cette définition moins courante. Notons n le nombre d'individus ayant répondu au questionnaire. On peut associer à chaque question un sous espace de l'espace $\mathbb{R}^n$ des fonctions numériques, définies sur l'ensemble des individus : le sous espace engendré par les variables indicatrices de ses réponses. (i.e. les colonnes du tableau disjonctif complet). Ce sous espace a pour dimensions le nombre de réponses proposées puisque les variables indicatrices des réponses à une même question sont orthogonales entre elles. C'est le sous espace des fonctions numériques ayant la même valeur pour les individus ayant choisi la même réponse. Les facteurs de l'analyse des correspondances peuvent être définies comme les facteurs de l'analyse multicanonique de ces sous espaces ; i.e. le premier facteur est le vecteur de $\mathbb{R}^n$ qui rend maximum la somme des cosinus carrés des angles entre ce vecteur et chacun des sous espaces, le second est orthogonal au premier et maximise le même critère, etc... Tous les sous espaces considérés contenant la droite des fonctions constantes, le premier facteur est le facteur constant, ou facteur trivial de l'analyse des correspondances ; les autres facteurs sont orthogonaux à ce facteur constant et sont donc des fonctions centrées. Ce point de vue enrichit l'interprétation des résultats, montre la cohérence de ce type d'analyse avec l'analyse en composantes principales normée qui est aussi une analyse multicanonique. Il permet aussi de proposer une projection des questions elles-mêmes sur les facteurs (cf. 6) qui facilite l'interprétation des résultats.

I.b - Le problème des "non réponses" ou des réponses très rares

Pour des raisons qui peuvent être très diverses, il est fréquent que des individus ne donnent pas de réponses à certaines questions. Si cette

"non réponse" traduit une attitude particulière, par exemple un refus volontaire, on peut la considérer comme une réponse particulière. On reste alors dans le cadre du tableau disjonctif complet. Mais, si cette non réponse n'a aucune signification particulière, n'est par exemple qu'un oubli, l'introduire en réponse supplémentaire risque de perturber les résultats. Du point de vue de l'analyse multicanonique, ceci revient à ajouter une dimension au sous espace associé à la question. Toutes les directions de ce sous espace jouant le même rôle dans la détermination des facteurs, cette direction non significative aura autant d'influence que les autres.

Le problème des réponses choisies très rarement est un peu analogue. Elles rendent les résultats instables, en ce sens que si l'on supprime les rares individus ayant choisi cette réponse, la dimension du sous espace associé à la question décroit de un, et les facteurs risquent d'être assez différents.

On peut supprimer du tableau de données ces réponses très rares, ou ne pas introduire les "non réponses". C'est une pratique courante mais le tableau obtenu, s'il est encore disjonctif n'est plus complet, il n'a donc plus toutes ses propriétés. En particulier, la marge sur l'ensemble des individus n'est plus constante. Cela modifie le profil des individus et si deux individus n'ont pas donné le même nombre de réponses, les modalités de réponses choisies simultanément par les deux individus augmentent leur distance. Ceci n'est guère logique.

Montrons le. Notons k_{ij} le général du tableau, $k_{i.}$ la marge sur l'ensemble I des individus, $k_{.j}$ la marge sur l'ensemble J des réponses et k l'effectif total du tableau.

$$k_{i.} = \sum_{j \in J} k_{ij} \qquad\qquad k_{.j} = \sum_{i \in I} k_{ij}$$

Rappelons la formule de la distance du χ^2 entre deux profils associés aux

individus i et i' :

$$D^2(i,i') = \sum_{j \in J} \left(\frac{k_{ij}}{k_{i.}} - \frac{k_{i'j}}{k_{i'.}} \right)^2 \frac{k}{k_{.j}}$$

si i et i' n'ont pas donné le même nombre de réponses, $k_{i.}$ et $k_{i'.}$ sont différents ; une réponse j choisie simultanément par les deux individus augmentera leur distance, puisque la différence $(k_{ij}/k_{i.} - k_{i'j}/k_{i'.})$ n'est pas nulle.

D'autre part, le tableau n'étant plus disjonctif complet, les propriétés agréables de l'analyse de ce type de tableau ne sont pas vérifiées. L'analyse ne peut être définie à partir des questions elles-mêmes, ce n'est plus une analyse multicanonique, on perd formellement une grande part de la cohérence de l'analyse et la double interprétation des facteurs.

I.c - La solution proposée. Propriétés générales

Nous proposons ici une solution dans laquelle les non réponses, -ou les réponses rares- sont supprimées sans les inconvénients que nous venons d'évoquer.

Dans cette solution, on garde la double interprétation de l'analyse des correspondances multiples : analyse d'un tableau croisant individus et réponses, et analyse multicanonique de sous espaces de $\mathbb{R}^n$ associés à chaque question.

Point de vue analyse des correspondances

Le tableau traité est le tableau disjonctif incomplet croisant les individus I et les réponses J prises en compte. La méthode appliquée est une variante de l'analyse des correspondances qui consiste à remplacer la marge sur I du tableau, qui n'est pas constante puisque le tableau est incomplet, par une marge constante. Les calculs et les propriétés des facteurs

sont tout à fait analogues à ceux de l'analyse des correspondances.

Nous présentons dans le paragraphe II cette variante dans un cadre un peu plus général que celui que nous utilisons ici, puisque nous remplaçons une des marges d'un tableau quelconque par une marge imposée quelconque. Nous y indiquons d'autres applications possibles de cette variante.

Voyons, dans le cas particulier qui nous intéresse, ce que deviennent les distances entre éléments de I et entre éléments de J lorsque la marge $k_{i.}/k$ est remplacée par la marge constante $1/n$. Dans la distance entre individus, la marge sur I intervient dans la définition du profil. En remplaçant cette marge par la marge constante $1/n$, le profil de l'individu i devient $k_{ij}(n/k)$ et la distance entre les profils associés aux individus i et i' :

$$D^2(i,i') = (n^2/k) \sum_{j \in J} (k_{ij} - k_{i'j})^2 (1/k_{.j})$$

L'illogisme que nous avions remarqué dans la distance classique du χ^2 disparaît ; seules les réponses différentes de i et i' augmentent cette distance. De plus, cette distance est tout à fait analogue à celle qui est utilisée dans les tableaux disjonctifs complets.

Pour l'ensemble J, le remplacement de $k_{i.}/k$ par $1/n$ ne modifie pas les profils des éléments, mais seulement leur distance :

Distance du χ^2 : $D^2(j,j') = \sum_{i \in I} (k_{ij}/k_{.j} - k_{ij'}/k_{.j'})^2 (k/k_{i.})$

Distance de la variante : $D^2(j,j') = n \sum_{i \in I} (k_{ij}/k_{.j} - k_{ij'}/k_{.j'})^2$

La distance entre j et j' est alors exactement celle que l'on obtiendrait dans le tableau disjonctif complet où les non réponses seraient introduites.

Les facteurs sur I et sur J seront, par définition, les projections, sur leurs axes d'inertie, de nuages de points respectant ces distances. Notons que le poids affecté aux individus étant fixé par la marge sur I, ils seront tous égaux.

Nous verrons au paragraphe II, que la dualité de l'analyse des nuages des individus et des réponses existe dans la variante proposée. Il y a donc des formules de transition des facteurs de l'un vers les facteurs de l'autre. D'où, une représentation simultanée s'interprétant à peu près comme celle de l'A.F.C.. Ces formules permettent aussi la projection d'éléments supplémentaires.

Nous aurons donc à peu près toutes les propriétés de l'analyse des correspondances, avec des distances entre individus et entre réponses légèrement différentes de celles de l'A.F.C., qui paraissent plus satisfaisantes dans ce cas particulier.

Point de vue analyse multicanonique

A chaque question, on associe encore le sous espace engendré par les variables indicatrices de ses réponses. Ce sous espace contient les fonctions qui prennent la même valeur pour les individus qui ont choisi la même réponse, et qui prennent la valeur 0 pour les individus, -s'il y en a- qui sont sans réponse.

Remarquons que les sous espaces associés aux questions avec non réponse ne contiennent pas la droite des constantes.

Une analyse multicanonique brutale de ces sous espaces donnerait des facteurs généralement non centrés qui optimiseraient un critère peu intéressant.

Dans le cas des tableaux disjonctifs complets, où chaque question

déterminait une partition de l'ensemble des individus, et où les facteurs de l'analyse multicanonique étaient centrés, le critère optimisé s'interprétait bien : le cosinus carré de l'angle entre un facteur et un sous espace associé à une question était le rapport de corrélation entre le facteur, (variable numérique centrée), et la question (variable qualitative). (cf. 6). C'était aussi le rapport entre l'inertie sur ce facteur des centres de gravité des classes d'individus ayant choisi la même réponse par l'inertie totale du facteur (cf. ci-dessous). Plus ce rapport d'inertie "inter" est proche de 1, plus les individus qui ont donné la même réponse sont groupés sur le facteur et plus le facteur représente bien la variable qualitative. Pour résumer plusieurs variables qualitatives, il était logique de choisir des fonctions numériques maximisant la somme de ces rapports.

Si le facteur n'est pas centré, ce cosinus carré ne peut plus s'interpréter ainsi.

La solution classique, pour obtenir des facteurs multicanoniques centrés serait de centrer les variables indicatrices des réponses, avant de construire les sous espaces. Cette solution ne convient pas ici, car elle est équivalente à la solution que nous voulions écarter, qui consistait à ajouter une réponse "non réponse" au tableau de données le rendant disjonctif complet.

Nous proposons d'imposer aux facteurs de l'analyse multicanonique d'être centrés sans modifier les sous espaces : nous chercherons donc un vecteur de $\mathbb{R}^n$, orthogonal à la droite des constantes, qui rend maximum la somme des cosinus carrés de ses angles avec les sous espaces considérés, un second orthogonal au premier, etc...

La solution de ce problème est donnée dans le paragraphe III dans un cadre un peu plus général : on imposera aux facteurs d'être ortho-

gonaux à un sous espace donné quelconque. C'est ce que nous appelons analyse multicanonique avec contrainte linéaire.

Ceci nous permet d'obtenir des facteurs centrés qui rendent maximum la somme des rapports d'inertie "inter". Le "rapport d'inertie inter" d'un facteur pour une question est le quotient de l'inertie des centres de gravité des classes d'individus ayant donné la même réponse par l'inertie totale du facteur. Plus ce rapport est proche de 1, plus les individus ayant donné la même réponse sont groupés et les sans réponses proches de zéro. On respecte donc tout à fait l'esprit de l'analyse des tableaux disjonctifs complets.

Démontrons maintenant cette affirmation. Pour cela, il suffit de démontrer que si V est un vecteur de $\mathbb{R}^n$, orthogonal à la droite des constantes et E_q le sous espace de $\mathbb{R}^n$ engendré par les variables indicatrices des réponses à la question q, alors $\cos^2(V, E_q)$ est égal au rapport "d'inertie inter de V pour q". Prenons V de norme -ou d'inertie- égale à 1.

Calculons d'abord le rapport d'inertie inter. Le centre de gravité sur V des individus ayant choisi la même réponse j à la question q est : $V(j) = \sum_i k_{ij} V(i)/k_{.j}$. Le rapport d'inertie inter est :

$$\sum_{j\in q} k_{.j} V^2(j) = \sum_{j\in q} (1/k_{.j}) \{\sum_i k_{ij} V(i)\}^2$$

Calculons maintenant $\cos^2(V,E_q)$. Les variables indicatrices e_j des modalités de réponses de q forment une base orthogonale de E_q. Donc

$$\cos^2(V,E_q) = \sum_{j\in q} \cos^2(V,e_j)$$

$$\text{Or} \quad \cos^2(V,e_j) = \{<V,e_j>/||e_j||\}^2$$

$$\{\sum_i k_{ij} V(i)\}^2/k_{.j}$$

Point de vue commun

Nous montrerons au paragraphe IV que les deux solutions proposées pour les questionnaires avec non réponses donnent les mêmes résultats.

Cette solution est donc très cohérente. Elle présente, comme l'analyse des correspondances multiples, tous les avantages d'une double interprétation de l'analyse : au niveau des réponses et au niveau des questions. Une application brutale des deux méthodes à ce type de tableau ne donnerait pas cette équivalence et en respecterait beaucoup moins l'esprit que les variantes proposées.

Il est évident que lorsque le tableau est complet, on obtient les résultats classiques.

Nous donnons au § IV les formules précises et les résultats particuliers obtenus dans le traitement proposé.

II - ANALYSE DES CORRESPONDANCES AVEC MARGE MODIFIEE.

Dans cette technique d'analyse factorielle tout à fait analogue à l'analyse des correspondances, l'une des marges du tableau de données est remplacée, partout où elle joue un rôle en Analyse des Correspondances par une marge imposée.

II.a - Exemples d'applications

II.a-1 Rappelons d'abord notre cas particulier des tableaux disjonctifs incomplets. Nous avons vu, que pour les individus comme pour les variables, la distance du χ^2 n'était pas très satisfaisante ; et que si l'on remplaçait dans chacune des deux formules de distances la marge sur les individus par une marge constante, les distances obtenues paraissaient plus logiques.

II.a-2 On traite quelquefois des sous tableaux d'un tableau de données. La métrique induite sur I dans l'analyse de $I \times J_1$ où J_1 est un sous ensemble de J est différente de celle qui est induite dans l'analyse de $I \times J$ si les marges sur I des deux tableaux sont différentes. Imposer la marge de $I \times J$ dans l'analyse du sous tableau permet de travailler avec la métrique induite par $I \times J$ et d'analyser le sous nuage $\mathscr{N}(J_1)$ de $\mathscr{N}(J)$. (cf. 10, 16).

II.a-3 Pour ajuster un tableau de fréquence à des marges imposées (cf. 14), nous proposons au § II.d d'appliquer la formule de reconstitution des données à des facteurs obtenus dans des analyses où les marges du tableau sont remplacées par les marges imposées.

II.b - Les calculs, les propriétés et les formules

II.b-1 Notation :

Soit I, J deux ensembles finis, un tableau rectangulaire sur

IxJ de nombres positifs ou nuls et de somme 1 est noté :

$$f_{IJ} = \{f_{ij} \ / i \in I, \ j \in J\}$$

Les marges de ce tableau sont notées f_I et f_J :

$$f_I = \{f_{i.} \ / i \in I\} \qquad f_{i.} = \sum_j \{f_{ij} \ / j \in J\}$$

$$f_J = \{f_{.j} \ / j \in J\} \qquad f_{.j} = \sum_i \{f_{ij} \ / i \in I\}$$

On note g_I , un ensemble indicé par I, de nombres positifs et de somme 1. Nous l'appellerons marge modifiée, ou imposée au tableau f_{IJ} .

II.b-2 Le nuage $\mathcal{N}(I)$

Dans l'analyse proposée le point i est représenté dans $\mathbb{R}_I$ par le profil $\tilde{f}{}^i_J = f_{iJ}/g_i$ muni du poids g_i . Profil et poids sont donc différents de ceux de l'Analyse des Correspondances. On note $\mathcal{N}(I)$ ce nuage.

La métrique de l'espace $\mathbb{R}_J$ est la même qu'en Analyse des Correspondances et le centre de gravité du nuage est aussi le même ; il a pour coordonnées $f_{.j}$.

Par définition, les facteurs sur I sont les projections des points du nuage $\mathcal{N}(I)$ sur ses axes principaux d'inertie. Pour calculer ces facteurs, on peut diagonaliser une matrice de dimension I ou les déduire des vecteurs propres d'une matrice de dimension J par une formule de projection. Nous donnons le terme général de ces matrices et la formule de projection dans le paragraphe II.b-4.

II.b-3 Le nuage $\mathcal{N}(J)$

Le point j est représenté, exactement comme en A.F.C. par le profil $f^j_I = f_{Ij}/f_{.j}$. Son poids est le même qu'en Analyse des Correspondances, c'est $f_{.j}$.

Par contre, la métrique définie sur l'espace ambiant $\mathbb{R}_I$ n'est pas la même, c'est la métrique diagonale $1/g_i$ et non pas $1/f_i$. Le centre de gravité du nuage est $f_{i.}$, comme en A.F.C. ; mais, pour garder une certaine cohérence, métrique et centre de gravité étant liés et représentant tous deux une certaine moyenne de référence, nous ferons l'analyse de ce nuage à partir du point de coordonnée g_i . Cela présente surtout le très grand avantage de conserver aux analyses de $\mathcal{N}(I)$ et de $\mathcal{N}(J)$ la dualité qui existe en Analyse des Correspondances. On aura donc des formules de transition permettant de passer des facteurs de l'un des nuages aux facteurs de l'autre et une représentation simultanée des deux ensembles.

II.b-4 Les formules

Le tableau ci-dessous résume les deux nuages étudiés (la marge réelle est $f_{i.}$ et la marge modifiée g_i) :

nuage	coordonnées du point	poids	c.d.g.	origine des axes	métrique	facteurs
$i \in \mathcal{N}(I)$	f_{iJ}/g_i	g_i	$f_{.j}$	$f_{.j}$	$1/f_{.j}$	$\mathcal{F}_s$
$j \in \mathcal{N}(J)$	$f_{Ij}/f_{.j}$	$f_{.j}$	$f_{i.} \neq$	g_i	$1/g_i$	$\mathcal{G}_s$

Pour calculer les facteurs, il faut diagonaliser des matrices. Rappelons les. (cf. 2, 12, 13). Notons X le tableau des coordonnées des points d'un nuage à partir de l'origine choisie, X' son transposé, P le tableau diagonal des poids affectés aux points et D la matrice du produit scalaire (métrique) de l'espace ambiant. Les facteurs sont les vecteurs propres de XDX'P. On peut les obtenir aussi en diagonalisant DX'PX et en appliquant aux vecteurs propres obtenus X.

Calculons le terme général de ces matrices. Pour le nuage $\mathcal{N}(I)$, notons M la matrice (I×I) dont les vecteurs propres sont les facteurs $\mathcal{F}_s$, et N la matrice (J×J) dont les vecteurs propres donnent $\mathcal{F}_s$ en leur appliquant X.

$$M_{ii'} = \sum_{j \in J} \frac{f_{ij}\, f_{i'j}}{g_i\, f_{.j}} + g_{i'} - f_{i'} - \frac{f_i\, g_{i'}}{g_i}$$

$$N_{jj'} = \sum_{i \in I} \frac{f_{ij}\, f_{ij'}}{f_{.j}\, g_i} - f_{.j'}$$

En calculant les matrices correspondantes pour le nuage $\mathcal{N}(J)$, on s'aperçoit que la matrice (J×J) dont les vecteurs propres sont les facteurs $\mathcal{G}_s$ est égale à N ; et que la matrice (I×I) dont les vecteurs propres donnent $\mathcal{G}_s$ en leur appliquant le tableau des coordonnées Y est égale à M.

Ceci montre la dualité de l'analyse des deux nuages. Les valeurs propres des deux analyses sont égales, $\mathcal{F}_s$ se déduit de $\mathcal{G}_s$ en appliquant X et $\mathcal{G}_s$ de $\mathcal{F}_s$ en appliquant Y. On a donc des formules de transition des uns vers les autres. Ces formules permettent aussi le traitement d'éléments supplémentaires car elles correspondent aux projections sur les axes d'inertie des nuages.

Remarquons que les facteurs $\mathcal{F}_s$ sont centrés pour la mesure g_i puisque l'analyse de $\mathcal{N}(I)$ se fait en prenant son centre de gravité comme origine :

$$\sum_i g_i\, \mathcal{F}_s(i) = 0$$

Par contre, les facteurs $\mathcal{G}_s$ ne le sont généralement pas :

$$\sum_j f_{.j}\, \mathcal{G}_s(j) \neq 0$$

Remarquons aussi que les $\mathcal{F}_s$ sont orthogonaux deux à deux pour

la métrique g_I et les $\mathcal{G}_s$ pour la métrique f_J .

$$\sum_i g_i \mathcal{F}_s(i) \mathcal{F}_{s'}(i) = \sum_j f_{.j} \mathcal{G}_s(j) \mathcal{G}_{s'}(j) = 0 \qquad \text{si} \quad s \neq s'$$

En effet la matrice M est symétrique pour la métrique g_I et N est symétrique pour f_J .

Ecrivons la formule de transition de $\mathcal{F}_s$ vers $\mathcal{G}_s$. On note λ_s la valeur propre associée, et on impose à $\mathcal{F}_s$ et $\mathcal{G}_s$ d'avoir pour norme $\sqrt{\lambda_s}$.

$$\mathcal{G}_s(j) = (1/\sqrt{\lambda_s}) \sum_i (f_{ij}/f_{.j} - g_i) \mathcal{F}_s(i)$$

$$= (1/\sqrt{\lambda_s}) \sum_i (f_{ij}/f_{.j}) \mathcal{F}_s(i)$$

Cette formule qui est exactement celle de l'Analyse des Correspondances s'est simplifiée car $\mathcal{F}_s$ est centré. Dans la représentation simultanée de I et de J, les éléments de J vérifieront donc la relation barycentrique classique.

Ecrivons maintenant la formule de transition de $\mathcal{G}_s$ vers $\mathcal{F}_s$.

$$\mathcal{F}_s(i) = (1/\sqrt{\lambda_s}) \sum_j (f_{ij}/g_i - f_{.j}) \mathcal{G}_s(j)$$

$$= (1/\sqrt{\lambda_s}) \{\sum_j (f_{ij}/g_i) \mathcal{G}_s(j) - \sum_j f_{.j} \mathcal{G}_s(j)\}$$

Il y a deux différences entre cette formule et celle de l'A.F.C. Dans le premier terme g_i intervient à la place de $f_{i.}$, ce n'est plus exactement la coordonnée du barycentre des $\mathcal{G}_s(j)$ affectés des poids f_{ij} . Le deuxième terme $\sum_j f_{.j} \mathcal{G}_s(j)$ est indépendant de i. C'est la projection du centre de gravité du nuage $\mathcal{N}(J)$, nous le calculerons systématiquement et le représenterons sur les facteurs. Il montre le décalage à faire subir aux éléments de J pour avoir une relation analogue à la relation barycentrique.

Les contributions absolues et relatives se calculent par les

formules habituelles. Nous indiquons ici la valeur de l'inertie des éléments i et j par rapport aux origines choisies.

$$\text{Inertie de } i = g_i \sum_i (f_{ij}/g_i - f_{.j})^2 (1/f_{.j})$$

$$\text{Inertie de } j = f_{.j} \sum_i (f_{ij}/f_{.j} - g_i)^2 (1/g_i)$$

Les calculs sont donc tout à fait analogues à ceux de l'analyse des correspondances, il suffit de faire quelques modifications au programme classique.

En écrivant les coordonnées des points de l'un des nuages dans la base de ses axes d'inertie, on obtient la formule de reconstitution des données :

$$f_{ij} = g_i \, f_{.j} \, (1 + \sum_s \mathcal{F}_s(i) \mathcal{G}_s(j)/\sqrt{\lambda_s})$$

II.c - Ajustement d'un tableau de fréquences à des marges

II.c-1 Le problème

Un exposé du problème et un panorama des techniques utilisées est donné dans l'article de J.L. MADRE (cf. 14).

Avec les notations du § II.b-1, on peut poser le problème dans les termes suivants : étant donné, d'une part, un tableau f_{IJ} de marges f_I et f_J et, d'autre part, deux ensembles g_I et g_J de nombres positifs et de somme 1, indicés par I et J ; on cherche à construire un tableau g_{IJ} de marge g_I et g_J "aussi proche que possible" du tableau g_{IJ} .

Dans l'article précité, deux nouvelles techniques étaient proposées basées sur la "formule de reconstitution des données" que nous rappelons ici :

(1) $$g_{ij} = g_i \, g_j \, (1 + \sum_s \mathcal{F}_s(i) \, \mathcal{G}_s(j)/\sqrt{\lambda_s})$$

Cette formule est exacte lorsque $\mathcal{F}_s$ et $\mathcal{G}_s$ sont les facteurs de l'A.F.C. du tableau g_{IJ} associés aux valeurs propres λ_s. Elle va permettre de construire un tableau g_{IJ} en l'appliquant à des fonctions $\mathcal{F}_s$ et $\mathcal{G}_s$ et des nombres λ_s calculés à partir du tableau f_{IJ}.

Pour que ce tableau g_{IJ} ait pour marges g_I et g_J, il suffit que les fonctions $\mathcal{F}_s$ soient centrées pour la loi g_I et les fonctions $\mathcal{G}_s$ centrées pour g_J. Si, de plus, les fonctions $\mathcal{F}_s$ (resp. $\mathcal{G}_s$) sont orthogonales deux à deux pour la loi g_I (resp. g_J), le tableau reconstruit g_{IJ} admet $\mathcal{F}_s$ et $\mathcal{G}_s$ comme facteurs de son analyse des correspondances.

Dans l'article en question, les fonctions $\mathcal{F}_s$ et $\mathcal{G}_s$ étaient calculées à partir des facteurs de l'A.F.C. du tableau f_{IJ}. Dans un cas on se contentait de centrer ces facteurs pour les lois g_I et g_J, dans l'autre on orthonormalisait leur suite.

II.c-2 Méthode proposée et propriétés du tableau reconstruit

Nous proposons d'introduire les lois g_I et g_J dans l'analyse du tableau f_{IJ} en appliquant l'analyse des correspondances avec marge modifiée. Les fonctions $\mathcal{F}_s$ et $\mathcal{G}_s$ sont calculées séparément. Les premiers en remplaçant la marge f_I par g_I, les seconds en remplaçant la marge f_J par g_J. La ressemblance entre le tableau reconstruit g_{IJ} et le tableau initial f_{IJ} se traduit précisément géométriquement, ce qui n'est pas le cas pour les autres méthodes. Le nuage $\mathcal{N}(I)$ représentant I dans l'A.F.C. du tableau g_{IJ} a :

a) les mêmes facteurs que le nuage représentant I dans l'A.F.C. de f_{IJ} avec la marge modifiée g_I. Les inerties de ces facteurs ne sont pas exactement les mêmes dans les deux analyses, mais sont très proches. Les distances entre les profils g_{iJ}/g_i sont donc, à très peu de choses près, celles des profils f_{iJ}/g_i et leur représentation par les facteurs sont identiques.

b) les mêmes axes d'inertie dans l'espace $\mathbb{R}_J$ que le nuage représentant I dans l'A.F.C. de f_{IJ} avec la marge modifiée g_J. Il a donc la même forme générale que le nuage des profils f_{iJ}/f_i si on prend dans $\mathbb{R}_J$, le point de coordonnée g_j comme origine et la métrique $1/g_j$.

Le nuage représentant J a les propriétés exactement symétriques.

II.c-3 Formule exacte et démonstration

Rappelons d'abord (cf. § II.b-4) que les facteurs sur I de l'analyse du tableau f_{IJ} avec la marge imposée g_I sont centrés et orthogonaux pour la loi g_I. Nous notons $\mathcal{F}_s$ ces facteurs et μ_s les valeurs propres associées.

De même, les facteurs sur J de l'analyse de f_{IJ} avec la marge imposée g_J sont centrés et orthogonaux pour la loi g_J. Notons $\mathcal{G}_s$ ces facteurs et ν_s les valeurs propres associées.

Pour construire g_{IJ}, nous appliquerons la formule (1) de II.c-1 aux fonctions $\mathcal{F}_s$ et $\mathcal{G}_s$ (ordonnées dans l'ordre décroissant de leur valeur propre).

Les fonctions $\mathcal{F}_s$ (resp. $\mathcal{G}_s$) étant centrées pour la loi g_I (resp. g_J), le tableau g_{IJ} a pour marge g_I et g_J.

Les fonctions $\mathcal{F}_s$ (resp. $\mathcal{G}_s$) étant orthogonales deux à deux pour la loi g_I (resp. g_J) les facteurs de l'A.F.C. du tableau g_{IJ} sont exactement les fonctions $\mathcal{F}_s$ et $\mathcal{G}_s$. Il faut encore choisir λ_s et il y a pour ces facteurs un petit problème de norme. Dans l'analyse de g_{IJ}, la valeur propre associée au couple $\mathcal{F}_s$, $\mathcal{G}_s$ est, si l'on prend λ_s quelconque : $\mu_s \nu_s/\lambda_s$. Pour que λ_s soit égal à cette valeur propre, on posera $\lambda_s = \sqrt{\mu_s \nu_s}$. Les facteurs de g_{IJ}, de norme $\sqrt{\lambda_s}$ sont donc : $\sqrt[4]{\nu_s/\mu_s}\,\mathcal{F}_s$ et $\sqrt[4]{\mu_s/\nu_s}\,\mathcal{G}_s$.

(Remarquons que les racines quatrièmes sont généralement proches de 1).

Ceci démontre la propriété indiquée en ⓐ au paragraphe précédent : les facteurs de l'analyse de g_{IJ} sont, à la norme près, les facteurs $\mathcal{F}_s$ (resp. $\mathcal{G}_s$) de l'analyse de f_{IJ} avec la marge g_I (resp. g_I).

Pour démontrer la propriété ⓑ, il faut utiliser la dualité de l'analyse des deux nuages représentant I et J. Cette dualité implique que les axes d'inertie du nuage représentant I ont pour image, par la métrique de l'espace $\mathbb{R}_J$, les facteurs de J. (cf. 1).

Nous avons vu que les facteurs $\mathcal{G}_s$ de l'analyse de f_{IJ} avec la marge modifiée g_J étaient égaux, à la norme près, aux facteurs sur J de g_{IJ}. D'autre part, les métriques considérées sur $\mathbb{R}_J$ sont, dans les deux analyses, $1/g_J$. Les facteurs sont égaux, les métriques sont les mêmes, les images des facteurs par cette métrique sont les mêmes. Les axes d'inertie des nuages représentant I dans les deux analyses ont les mêmes directions.

II.c-4 Comparaison des tableaux g_{IJ} et f_{IJ}

En résumé, pour obtenir le tableau ajusté g_{IJ}, on fera deux analyses de f_{IJ} en imposant une fois la marge g_I et l'autre fois la marge g_J.

On appliquera ensuite la formule de reconstitution des données, éventuellement limitée aux premiers facteurs, aux facteurs sur I de la première analyse et aux facteurs sur J de la seconde pour construire g_{IJ}.

Nous avons vu les analogies entre les nuages associés au tableau g_{IJ} et des nuages associés au tableau f_{IJ} avec l'une des marges g_I ou g_J . Le tableau ci-dessous les résume pour les nuages associés à I, la situation est absolument symétrique pour les nuages associés à J.

Nuages représentant I

Tableau	Marge imposée	Coordonnées	Métrique	Axes	Facteurs
g_{IJ}		g_{iJ}/g_i	$1/g_J$	a_s	$\mathcal{F}_s$
f_{IJ}	g_I	f_{iJ}/g_i	$1/f_J$		$\mathcal{F}_s$
f_{IJ}	g_J	f_{iJ}/f_i	$1/g_J$	a_s	
f_{IJ}		f_{iJ}/f_i	$1/f_J$		

Mais il peut être intéressant de comparer ces deux tableaux sans introduire pour f_{IJ} une des marges imposées.

Pour comparer les profils des tableaux f_{IJ} et g_{IJ}, il est courant de faire l'analyse de g_{IJ} et de mettre en éléments supplémentaires le tableau f_{IJ} en lignes et en colonnes.

Or, les facteurs sur J de l'analyse de f_{IJ} avec la marge g_I -qui n'ont pas été utilisés dans la formule de construction de g_{IJ}- donnent exactement les projections des profils des colonnes de f_{IJ}. On peut le vérifier sur la formule de transition.

De même les facteurs sur I de l'analyse de f_{IJ} avec la marge g_J donnent les projections des profils des lignes de f_{IJ}.

On conservera donc les facteurs des deux analyses nécessaires à la construction de g_{IJ}. On aura ainsi, sans aucun calcul supplémentaire, l'analyse de g_{IJ} et les projections en éléments supplémentaires des profils des lignes et des données de f_{IJ} sur les facteurs de cette analyse.

II.d - L'analyse par sous tableaux

L'analyse des correspondances par sous tableaux (cf. 8, 10) est

une technique permettant de calculer, avec une bonne approximation, les résultats de l'analyse des correspondances d'un très grand tableau de données. L'idée est de diviser ce tableau en sous tableaux, de calculer les facteurs de ces sous tableaux, puis de calculer les résultats approchés du tableau entier à partir des premiers facteurs de tous les sous tableaux.

Notons I et J les deux ensembles en correspondance, et supposons que la division en sous tableaux ait été obtenue par une partition de J en r sous ensembles $J_1,\ldots,J_r$. Généralement les marges sur I du tableau entier IxJ et des sous tableaux sont différentes. (Ceci n'est pas le cas lorsque tous les sous tableaux sont disjonctifs complets).

Deux cas peuvent se présenter. Soit la partition de J a été obtenue en regroupant des éléments proches entre eux. Dans ce cas, chaque sous tableau présente une certaine unité, et les sous tableaux diffèrent entre eux. Généralement, leur marge sur I qui donne les coordonnées des centres de gravité des sous nuages, sont assez différentes.

Soit la partition de J a été obtenue en divisant J au hasard ou par échantillonnage. Alors les sous tableaux, ne présentent aucune unité, ne diffèrent guère entre eux a priori. Leurs marges sur I sont assez proches et ne présentent aucun intérêt particulier. Notons aussi, que dans certains cas, les sous tableaux peuvent être homogènes sans que leurs marges diffèrent. (Tableau disjonctif complet ou presque).

Ces remarques amènent à proposer plusieurs variantes de la méthode des sous tableaux.

Lorsque les marges des sous tableaux sont assez proches, et non significatives, le plus simple est de procéder à l'analyse de chaque sous tableau en imposant la marge du tableau entier. Géométriquement, ceci revient à analyser chaque sous nuage $\mathcal{N}_{(J_1)},\ldots,\mathcal{N}_{(J_2)}$ avec la métrique induite par

par le nuage entier $\mathcal{N}(J)$ en prenant comme origine le centre de gravité de $\mathcal{N}(J)$. (cf.

Par contre, lorsque les marges sont très différentes et significatives pour chaque tableau, il est bien préférable de faire l'analyse de ces sous tableaux avec leur marge propre. Dans ce cas on est amené à faire l'analyse du tableau des marges des sous tableaux, et à utiliser ses facteurs pour le calcul des facteurs approchés du tableau entier. Les formules du calcul des facteurs approchés sont un peu plus compliquées, mais les résultats de l'approximation sont bien meilleurs.

II.e - Traitement simultané de variables qualitatives et quantitatives

Nous avons proposé en (9) une analyse factorielle traitant simultanément variables qualitatives et variables quantitatives. Quand elles sont seules, les premières codées par des tableaux disjonctifs complets sont traitées par l'analyse des correspondances ; les secondes sont traitées par l'analyse en composantes principales. Or, l'analyse en composantes principales normées est, comme l'analyse des tableaux disjonctifs complets, une analyse multicanonique : celle des sous espaces de dimension 1 engendrés par les vecteurs de $\mathbb{R}^n$ représentant les variables centrées. Pour traiter simultanément les deux types de variables, nous avions proposé de faire une analyse multicanonique de l'ensemble des sous espaces de dimension 1 associés aux variables quantitatives et des sous espaces de dimension supérieure associés aux variables qualitatives. Un artifice de codage permet d'utiliser les programmes classiques d'analyse des correspondances : les variables qualitatives sont codées par le tableau disjonctif complet, les variables numériques par 2 colonnes ; l'une contient les valeurs $(1+x_i)/2$ où x_i est la valeur de la variable centrée normée pour l'individu i, l'autre colonne contient $(1-x_i)/2$.

Le codage par deux colonnes contenant une information redondante était nécessaire pour avoir une marge constante sur I, mais évidemment il augmente la taille du tableau et le temps de calcul.

On peut éviter ce double codage en utilisant, au lieu du programme classique d'analyse des correspondances la variante proposée en imposant une marge constante. Dans $\mathbb{R}^n$ la métrique et le centre de gravité du nuage sont les mêmes qu'avec le double codage. Les points représentant les deux colonnes étaient symétriques par rapport au centre de gravité et avaient le même poids. Pour obtenir la même inertie avec un seul point il suffit de lui donner un poids double. Nous coderons donc la variable quantitative par la colonne $1+x_i$.

III - ANALYSE MULTICANONIQUE AVEC CONTRAINTE.

III.a - Rappel sur l'analyse multicanonique

Rappelons que l'analyse multicanonique de p sous espaces $E_1,\ldots,E_p$ d'un même espace euclidien $\mathbb{R}^n$, est la recherche d'une suite de vecteurs orthogonaux rendant maximum le critère (cf. 3, 15, 4).

$$\sum_{p=1}^{P} \cos^2(\theta_p)$$

où θ_p est l'angle entre le p-ième sous espace et le vecteur cherché.

Rappelons aussi que ces vecteurs sont appelés facteurs de l'analyse multicanonique des sous espaces E_p et qu'ils sont obtenus en diagonalisant l'opérateur A :

$$A = \sum_{p=1}^{P} \Pi_p$$

où Π_p est l'opérateur de projection orthogonale sur E_p. En effet, un vecteur propre v de norme 1 de la matrice symétrique positive A, associé à sa plus grande valeur propre rend maximum v'Av, où v' est le transposé de A.

$$\text{Or, } v'Av = v'\left(\sum_p \Pi_p\right) v = \sum_p (v' \Pi_p v).$$

Si v est de norme 1, sa projection sur E_p, Π_p v a pour norme $\cos\theta_p$, et le produit scalaire de Π_p v par v, v' Π_p v vaut $\cos^2\theta_p$.

Le vecteur propre v est donc le premier facteur de l'analyse multicanonique, et la valeur propre associée vaut $\sum_p \cos^2\theta_p$. Les facteurs suivants sont les autres vecteurs propres de A ordonnés dans le sens décroissant des valeurs propres ; A étant symétrique, ses vecteurs propres sont orthogonaux.

III.b - L'analyse multicanonique avec contrainte

L'analyse multicanonique avec contrainte de p sous espaces est une analyse multicanonique où l'on impose aux facteurs d'être orthogonaux à un sous espace F.

Pour le traitement des questionnaires avec non réponse, le sous espace F est la droite des constantes, i.e. nous imposons seulement aux facteurs d'être centrés.

Le cas où F est un sous espace quelconque pouvant avoir d'autres applications, et la solution étant aussi simple, c'est celui que nous traitons ici.

Proposition

Ayant choisi une base orthonormée de l'espace ambiant, les facteurs de l'analyse multicanonique avec contrainte sont les vecteurs propres de la matrice symétrique :

$$A - YY'A - AYY' + YY'AYY'$$

où

- A *est la matrice de l'opérateur* $\sum_p \Pi_p$
- Y *est une matrice dont les r colonnes forment une base orthonormée de* F
- Y' *est la transposée de* Y

Les valeurs propres associées valent $\sum \cos^2\theta_p$ *, ces facteurs sont donc ordonnés dans l'ordre décroissant des valeurs propres.*

Démonstration

Soit v, un vecteur de norme 1.

Nous avons vu que $v'Av = \sum_p \cos^2\theta_p$.

Si on note $y_1,\ldots,y_R$ les R colonnes de Y, i.e. une base ortho-

normée de F, la contrainte "v orthogonal à F" est équivalente aux R contraintes :

$$y_1' v = 0, \ldots\ldots, y_R' v = 0$$

Le problème est donc de trouver v qui rende maximum v'Av sous ces R contraintes, plus la contrainte v'v = 1 qui impose à v d'avoir une norme 1.

Soient λ et μ_r $(1 \leqslant r \leqslant R)$ des multiplicateurs de Lagrange, la dérivation de la quantité

$$v'Av - \lambda(v'v-1) - \sum_{r=1}^{R} \mu_r v'y_r$$

par rapport aux différentes composantes de v, puis l'annulation de ces dérivées conduisent à la relation :

$$Av - \lambda v - \sum_{r=1}^{R} \mu_r y_r = 0 \qquad (1)$$

La solution v, vérifie $y_r' v = 0$ pour tout r compris entre 1 et R. Si on multiplie (1) à gauche par y_r', on obtient : $y_r' Av - \mu_r = 0$. D'où $\mu_r = y_r' Av$. La relation (1) devient :

$$Av - \lambda v - \sum_{r=1}^{R} y_r y_r' Av = 0$$

ou

$$Av - \lambda v - YY' Av = 0$$

ou encore

$$(A - YY'A)v = \lambda v$$

Donc la solution v est vecteur propre de A - YY'A. Mais cette matrice n'est pas symétrique, ce qui pose quelques problèmes pour l'obtenir. Nous les résolvons en construisant une matrice M symétrique qui a les mêmes valeurs propres et les mêmes vecteurs propres. Cette matrice M se déduit de A - YY'A en ajoutant 2 termes.

$$M = A - YY'A - AYY' + YY'AYY'$$

Démontrons les propriétés de M indiquées ci-dessus.

Il est évident que M est symétrique.

Montrons d'abord que pour tout vecteur propre v de M, $Y'v = 0$, i.e. v est orthogonal à F. En effet :

si $Mv = \lambda v$

$$Y'v = (1/\lambda)\ Y'\ Mv$$

$$= (1/\lambda)(Y'A - \{Y'Y\}\ Y'A - Y'AYY' + \{Y'Y\}\ Y'AYY')$$

$$= 0 \qquad \text{car } Y'Y \text{ est la matrice identité}$$

Or, pour tout vecteur v tel que $Y'v = 0$, il est évident que $Mv = (A-YY'A)v$ et réciproquement, Donc, la matrice M a les mêmes vecteurs propres et les mêmes valeurs propres que $A - YY'A$.

Pour terminer la démonstration de la proposition, il faut montrer que si v est un vecteur propre de M, de norme 1, associé à la valeur propre λ, cette valeur propre vaut $\sum_p \cos^2\theta_p$. Nous avons vu que $v'Av = \sum_p \cos^2\theta_p$. Or, si v est vecteur propre de M, il est vecteur propre de $A - YY'A$, et on aura :

$$Av - YY'Av = \lambda v$$

D'où $$v'Av = v'YY'Av + \lambda v'v$$

$$= \lambda$$

IV - APPLICATION DES DEUX METHODES AUX QUESTIONNAIRES AVEC NON REPONSE

Dans ce paragraphe nous donnons les formules des deux méthodes appliquées à ce cas particulier et nous démontrons que leurs résultats sont identiques.

Pour compléter ces résultats, nous proposons une représentation des questions elles-mêmes sur les facteurs ; l'analyse des correspondances avec marge constante ne donnant de représentation que des modalités des questions. Cette représentation, analogue à celle que nous avions proposé en (6) pour les tableaux disjonctifs complets met en évidence les liaisons entre les questions et les facteurs. Elle est basée sur l'interprétation "analyse multicanonique" des facteurs.

Rappelons d'abord que I note l'ensemble des n individus, J l'ensemble des réponses, k_{ij} le terme général (0 ou 1) du tableau disjonctif incomplet, $k_{i.}$ la somme de la ligne i et $k_{.j}$ celle de la colonne j, k l'effectif total.

IV.a - Analyse des correspondances avec marge imposée constante

Nous appliquons cette méthode en remplaçant la marge $k_{i.}/k$ par la marge constante 1/n. Les distances entre individus et entre réponses sont alors très proches de celles des tableaux disjonctifs complets.

Donnons le terme général des matrices M et N dont les vecteurs propres sont les facteurs de cette analyse.

$$M_{ii'} = (n/k) \sum_{j \in J} \{k_{ij} k_{i'j} / k_{.j}\} + 1/n - k_{i'}/k - k_{i}/k$$

$$N_{jj'} = (n/k) \sum_{i \in I} \{k_{ij} k_{ij'} / k_{.j}\} - k_{.j'}/k$$

Pour le calcul de ces facteurs, on procède comme en analyse des correspondances en diagonalisant une matrice symétrique, mais ici il n'y a pas de facteur trivial à supprimer.

Remarque :

On sait (cf. 1, 2, 12, 13) que l'analyse d'un tableau disjonctif complet est équivalente à celle du tableau de fréquence des modalités croisées deux à deux, qu'on appelle généralement tableau de Burt. Cette propriété n'est pas vérifiée ici. Par contre, il est clair que N se déduit facilement de ce tableau de fréquence. On pourra donc envisager de construire N à partir d'un tableau de données condensées, avec une seule colonne par question et l'indication de la modalité choisie comme dans le programme d'analyse des correspondances multiples de [12].

Voyons maintenant les formules de transition. On note $\mathcal{F}_s$, $\mathcal{G}_s$ le couple de facteurs d'ordre s associé à la valeur propre λ_s. Le facteur $\mathcal{F}_s$ est centré pour la mesure uniforme, $\sum_i \mathcal{F}_s(i) = 0$.

$$\mathcal{G}_s(j) = (1/\sqrt{\lambda_s}) \sum_i (k_{ij}/k_{.j}) \mathcal{F}_s(i)$$

C'est la formule barycentrique classique, une modalité de réponse est située, à $1/\sqrt{\lambda_s}$ près, au centre de gravité des individus qui l'on choisie.

L'autre formule est :

$$\mathcal{F}_s(i) = (1/\sqrt{\lambda_s})\{\sum_j (nk_{ij}/k) \mathcal{G}_s(j) - \sum_j (k_{.j}/k) \mathcal{G}_s(j)\}$$

le terme $\sum_j (k_{.j}/k) \mathcal{G}_s(j)$ représente la projection du centre de gravité des modalités de réponses étudiées. Nous l'indiquerons sur les graphiques. Par rapport à la relation barycentrique, tous les individus sont décalés par ce même terme.

On pourra mettre en éléments supplémentaires les "non réponses". Remarquons que, en considérant aussi les non réponses, le facteur $\mathcal{G}_s$ est centré puisque l'origine des axes est au centre de gravité de l'ensemble des réponses et non réponses.

L'interprétation des contributions absolues et relatives est exactement celle de l'Analyse des Correspondances classiques.

IV.b - Analyse multicanonique avec contrainte

Rappelons que nous avions proposé dans le §I.c d'associer à chaque question le sous espace de $\mathbb{R}^n$ engendré par les variables indicatrices de ses modalités de réponses, i.e. le sous espace des fonctions prenant la même valeur pour les individus ayant choisi la même réponse et prenant la valeur 0 pour les individus sans réponse.

Nous proposons de faire l'analyse multicanonique de ces sous espaces en imposant aux facteurs d'être centrés, i.e. orthogonal à la droite des constantes. Ceci, afin d'obtenir des facteurs centrés $\mathcal{F}_s$ qui rendent maximum la somme des "inerties inter" des questions : à une question, on associe les centres de gravité sur $\mathcal{F}_s$ des classes de population ayant choisi la même réponse, l'inertie inter de la question est l'inertie de ces centres de gravité.

Egalité des facteurs. Point de vue matriciel

Calculons le terme général de la matrice symétrique du § II.b que l'on diagonalise pour obtenir les facteurs de cette analyse multicanonique. Nous verrons qu'elle est égale à la matrice M du paragraphe précédent, ce qui implique l'égalité des facteurs.

Les variables indicatrices e_j d'une même question q étant orthogonales entre elles, la projection sur le sous espace E_q qu'elles engendrent

s'écrit facilement. Notons X_q le tableau ayant pour colonnes ces variables. C'est un sous tableau du tableau de données k_{ij}. Notons Δ_q une matrice carrée, de dimension égale au nombre de modalités de q, ayant sur la diagonale $1/k_{.j}$ (c'est l'inverse du carré de la norme de e_j) et 0 ailleurs. La projection sur E_q est $X_q \Delta_q X'_q$. La somme de ces projections est $A = \sum_q X_q \Delta_q X'_q$ son terme général est $a_{ii'} = \sum_{j \in J} k_{ij} k_{i'j} / k_{.j}$. La matrice notée Y au § II.b se réduit au vecteur unitaire de la droite des constantes de $\mathbb{R}^n$. Un calcul simple montre l'égalité de A - YY'A - AYY' + YY'AYY' et de M.

Egalité des facteurs. Point de vue géométrique

Un raisonnement plus géométrique peut être utilisé pour cette démonstration.

Dans l'analyse des correspondances avec la marge 1/n, une modalité j est représentée par le point de coordonnées $k_{Ij}/k_{.j}$. Son inertie par rapport à l'origine vaut n/k, quel que soit j : $(k_{.j}/k) \sum_i (k_{ij}/k_{.j})^2 \, n = n/k$.

D'autre part, le vecteur joignant l'origine à ce point est colinéaire à e_j. L'inertie de la projection de j sur un vecteur V vaut donc $(n/k) \cos^2(V, e_j)$.

La somme des inerties des projections de toutes les modalités de la variable q sur V vaut donc $(n/k) \cos^2(V, E_q)$.

La métrique sur $\mathbb{R}^n$ étant la métrique identité à (1/n) près, les facteurs $\mathcal{F}_s$ et les axes d'inertie du nuage des modalités sont homothétiques (cf. II.c-3). Ils sont orthogonaux à la droite des constantes, et orthogonaux entre eux. Ils rendent maximum la somme des inerties des projections de toutes les modalités de réponses, donc la somme des $\cos^2(V, E_q)$. Ce qui termine la démonstration.

IV.c - Projection des questions sur les facteurs

Nous avions proposé en (cf. 6), comme aide à l'interprétation des résultats d'une analyse des correspondances sur tableau disjonctif complet, de projeter les variables (ou questions) sur les facteurs. La double interprétation des facteurs que nous avons définis pour les tableaux incomplets permet de procéder de la même façon.

Nous avons vu au paragraphe I.c que pour un facteur $\mathcal{F}$ et une question q, $\cos^2(\mathcal{F},E_q)$ était le rapport de variance "inter" de $\mathcal{F}$ pour q. C'est donc une mesure du lien entre $\mathcal{F}$ et q.

Nous venons de voir qu'à (n/k) près, c'est aussi la contribution absolue des modalités de q au facteur $\mathcal{F}$. C'est donc, de ce point de vue encore, une mesure du lien entre $\mathcal{F}$ et q, et il se calcule très facilement.

Nous proposons de représenter q sur les facteurs en lui donnant comme coordonnées ces cosinus carrés. On mettra ainsi en évidence les liens entre un facteur et une question.

Pour cette représentation, nous parlons de projection car elle s'interprète ainsi dans l'espace des opérateurs symétriques de $\mathbb{R}^n$. C'est la projection d'un opérateur associé à la question sur des opérateurs associés aux facteurs.

Ces opérateurs sont des opérateurs de projection orthogonale.Pour un facteur, c'est la projection sur le sous espace qu'il engendre ; pour une question q c'est la projection sur le sous espace E_q .

Rappelons que l'espace des opérateurs symétriques est muni d'un produit scalaire : $\langle A,B\rangle$ = Trace AB .

Rappelons aussi que les opérateurs associés à deux facteurs distincts sont orthogonaux puisque les facteurs sont orthogonaux ; que la norme de ces opérateurs vaut un, puisque le sous espace engendré par un facteur est

de dimension un. Ils forment donc une base orthonormée incomplète de l'espace des opérateurs.

La représentation que nous proposons est une projection sur le sous espace engendré par les opérateurs associés aux premiers facteurs. Si deux questions sont bien représentées par leur projection, on mettra en évidence leur proximité ou leur distance dans l'espace des opérateurs ; on mettra donc en évidence la liaison entre ces deux questions.

IV.d - Méthode classique et variante proposée : comparaison

Si le nombre de réponses manquantes est très faible, ou bien réparti sur les individus, autrement dit si la marge sur I du tableau disjonctif incomplet est presque constante, les facteurs obtenus par l'analyse des correspondances classique et par la variante que nous proposons différeront très peu. En effet, le centre de gravité du nuage des modalités sera très proche de l'origine que nous imposons et la métrique du χ^2 très proche de l'identité.

Dans ce cas, on pourra négliger l'écart entre les résultats des deux méthodes et interpréter l'analyse classique avec les points de vue indiqués ci-dessus.

Nous donnons ci-dessous des bornes précises des écarts entre les facteurs des deux analyses. Ces bornes sont obtenues en comparant les matrices diagonalisées et en appliquant les résultats obtenus dans [7]. Elles montrent, ce qui est naturel, que plus la marge est proche de la marge constante et plus les valeurs propres de l'analyse sont bien séparées les unes des autres, plus les facteurs des deux analyses sont semblables.

Notons M la matrice dont les vecteurs propres sont les facteurs sur I de l'analyse avec marge constante (cf. § IV.a) et M' son homologue de l'analyse des correspondances classiques. On peut écrire :

$$M = DM' + R$$

où D est une matrice diagonale $\delta_i^{i'}\ n\ k_i/k$

et R une matrice symétrique de terme général $1/n + (k_i+k_{i'})/k+n\ k_i k_{i'}/k^2$

Notons θ l'angle dans R^J les facteurs de même ordre s des deux analyses. Notons λ_s la valeur propre associée à ce facteur dans l'analyse classique, ε l'écart minimum entre λ_s et les autres valeurs propres. Les théorèmes donnés en [7] permettent de majorer θ en fonction des plus grandes valeurs propres de D et de R, ce qui donne les inégalités suivantes :

$$\theta < \theta_1 + \theta_2$$

et

$$\sin 2\theta_1 \leqslant (\lambda_s/\varepsilon)(\sup_i k_i/\inf_i k_i - 1) \qquad \text{(cf. 7, chap. I - IV)}$$

$$\sin 2\theta_2 \leqslant 2\{1 - n \sum_i (k_i/k)^2\}/\varepsilon \qquad \text{(cf. 7, chap. I - II)}$$

V - CONCLUSION.

L'analyse des correspondances avec marge modifiée est une méthode facile à programmer, et d'utilisation simple puisque la technique et les résultats sont extrêmement proches de l'Analyse des Correspondances classique.

Nous avons proposé dans cet article plusieurs applications de cette méthode ; le traitement des questionnaires avec réponses manquantes ; l'ajustement d'un tableau de fréquence à des marges, une méthode d'analyse par sous tableaux, le traitement simultané de variables qualitatives et quantitatives. La liste n'est sans doute pas terminée et cette méthode parait donc avoir une place dans l'arsenal des méthodes d'analyse des données.

Pour l'analyse multicanonique avec contrainte, nous n'avons jusqu'ici comme application que les tableaux disjonctifs incomplets. Comme dans ce cas particulier elle est équivalente à l'analyse des correspondances avec marge modifiée, elle sert alors essentiellement à montrer la cohérence de cette méthode et à compléter l'interprétation de ses résultats.

- BIBLIOGRAPHIE -

[1] BENZECRI et Collaborateurs : *"L'analyse des données"* - Dunod 1973.

[2] CAILLEZ F., PAGES J.P. : *"Introduction à l'analyse des données"* - SMASH 1976.

[3] CARROLL J.P. : *"A generalisation of canonical correlation analysis to three or more sets of variable"* - Proc. 76th Amer. Psych. Assoc. 1968.

[4] DAUXOIS J., POUSSE M. : *"Les analyses factorielles en calcul des probabilités et en statistique. Essai d'étude synthétique"* - Thèse Université Paul Sabatier, Toulouse 1976.

[5] DIDAY E. et Collaborateurs : *"Optimisation en classification automatique"*.

[6] ESCOFIER B. : *"Une représentation des variables dans l'analyse des correspondances multiples"* - Revue de Statistique Appliquée volume XXVII, n° 4, 1979.

[7] ESCOFIER B. : *"Stabilité et approximation en Analyse Factorielle"* - Thèse, Paris 6, 1979.

[8] ESCOFIER B. : *"Analyse des correspondances par sous tableaux"* - Rapport IRISA 1979.

[9] ESCOFIER B. : *"Traitement simultané des variables qualitatives et quantitatives en Analyse Factorielle"*.

[10] ESCOFIER B. : *"Analyse par sous tableaux - Méthode n° 3"* - Support du cours "Développements récents en Analyse des Données" Mars 1980, INRIA.

[11] ESCOUFIER Y. : *"Opérateur associé à un tableau de données"* - Annales de l'INSEE, n° 22-23, 1976.

[12] LEBART L., MORINEAU A., TABARD N. : *"Technique de la description statistique"* - Dunod 1977.

[13] LEBART L., MORINEAU A., FENELON J.P. : *"Traitement des données statistiques"* - Dunod 1979.

[14] MADRE J.L. : *"Méthodes d'ajustement d'un tableau à des marges"* - Cahiers de l'Analyse des données, volume V, n° 2, 1980.

[15] SAPORTA G. : *"Liaison entre plusieurs ensembles de variables et codage de données qualitatives"* - Thèse de 3ème cycle, Paris 6.

[16] OK Y. : *"Analyse factorielle typologique"* - Thèse de 3ème cycle, Paris 6, 1975.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique